Chapter 1
Quality Management in Web Warehouses

Adriana Marotta
Universidad de la República, Uruguay

Laura González
Universidad de la República, Uruguay

Lorena Etcheverry
Universidad de la República, Uruguay

Bruno Rienzi
Universidad de la República, Uruguay

Raúl Ruggia
Universidad de la República, Uruguay

Flavia Serra
Universidad de la República, Uruguay

Elena Martirena
Universidad de la República, Uruguay

ABSTRACT

Web Warehouses (WW) are data warehouses that consolidate data from the Web. The process of building them presents several challenges, most of them related to the autonomy and dynamicity of Web sources. In this context, managing quality aspects becomes a fundamental issue since information about quality is needed to properly select Web sources to populate the WW. Additionally, measuring and propagating quality values to the WW might provide final users with valuable information to improve decision-making processes. In this chapter, we present a reference architecture for quality aware Web Warehouses, which specifies the main components to evaluate and manage quality aspects through all the life cycle of a WW and considers quality regarding data and services.

DOI: 10.4018/978-1-61350-038-5.ch001

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

Supported by new technology trends, like Web 2.0, Cloud Computing and Web Services, information available on the Web is increasing every day. Consequently, Web Warehouses (WW), Data Warehouses (DW) which consolidate data from the Web (Cheng et al., 2000; Marotta et al., 2002; Soper, 2005), have become a valuable tool for decision making in many areas.

However, given the dynamic and autonomous nature of Web Data Sources (WDSs) (Bhowmick et al., 2004), the process of building a WW presents major challenges. First, web data can be delivered through many heterogeneous formats and protocols, including among others HTML pages, XML documents, RSS or ATOM feeds, SOAP Web Services and Restful Web Services. Second, WDSs are usually managed by third-parties, leading to uncertainty in terms of availability, cost and unexpected changes in the format or protocols they use to deliver data.

In this context, managing quality aspects becomes a fundamental issue. On the one hand, quality information is needed to properly select the WDSs to populate the WW. Additionally, measuring and propagating quality values to the WW might provide end users with valuable information to improve decision-making processes.

This chapter presents a Reference Architecture for Quality Aware Web Warehouses, which specifies the main components to evaluate and manage quality aspects through all the lifecycle of a WW. In this reference architecture, quality is considered regarding data and services. Data Quality (DQ) deals, on one hand, with all the quality aspects related with web data, like completeness, freshness and accuracy, and on the other hand, with quality aspects that are of particular interest in the context of Data Warehouses, like hierarchy completeness and measure precision. Quality of Service (QoS) deals with the aspects concerning the services that provide or manipulate the data, like availability, response time and reputation. Based on this quality information, it is possible to perform a quality aware discovery of WDSs, and to provide a quality driven runtime adaptation mechanism, during the extraction stage, to deal with the highly dynamic nature of the Web. As well, the platform addresses the measurement, propagation, aggregation and management of quality metadata, through all the stages needed to populate a WW. This provides the end user valuable quality information, which can be leveraged to take more accurate and confident decisions.

In order to manage quality information we take the approach followed in (Etcheverry et al., 2008), where quality is characterized via multiple dimensions, each of which captures a high-level aspect of quality. Each quality dimension consists of a set of quality factors, where each factor represents a particular aspect of a quality dimension. Finally, quality metrics are instruments to measure a particular quality factor. Several metrics might exist for the same quality factor.

Our proposal mainly focuses on three aspects: (i) a reference architecture for managing quality in a WW, (ii) management of data and service quality in web data sources, and (iii) management of data quality in the DW component of the WW. These three aspects are closely related. While the last two items deal with the management of data and service quality, the reference architecture provides the global environment to connect the components that perform the different involved tasks, including quality management.

This chapter is organized as follows. Next section presents existing works covering some of the topics we address within this chapter and how our solution is positioned with respect to them. Then, we present and describe a reference architecture for quality aware WW, focusing on how quality is managed throughout the architecture. After that, we present in detail a proposal for managing web-data quality and a proposal for managing DW quality, describing the development of a case study in which we exemplify the main stages in a WW implementation following our approach. Finally, we identify future research directions and we present the conclusions of the chapter.

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/quality-management-web-warehouses/58409

Related Content

Modelling in Support of Decision Making in Business Intelligence
Roumiana Ilieva, Malinka Ivanova, Tzvetilina Peychevaand Yoto Nikolov (2021). *Integration Challenges for Analytics, Business Intelligence, and Data Mining (pp. 115-144).*
www.irma-international.org/chapter/modelling-in-support-of-decision-making-in-business-intelligence/267869

Taxonomy Outline of Big Data Analytics Literature
Sapna Sinha, Vishal Bhatnagarand Abhay Bansal (2014). *Encyclopedia of Business Analytics and Optimization (pp. 2456-2471).*
www.irma-international.org/chapter/taxonomy-outline-of-big-data-analytics-literature/107427

Critical Barriers to Business Intelligence Open Source Software Adoption
Placide Poba-Nzaou, Sylvestre Uwizeyemunguand Mariem Saada (2019). *International Journal of Business Intelligence Research (pp. 59-79).*
www.irma-international.org/article/critical-barriers-to-business-intelligence-open-source-software-adoption/219343

The Prediction of Workplace Turnover Using Machine Learning Technique
Youngkeun Choiand Jae Won Choi (2021). *International Journal of Business Analytics (pp. 1-10).*
www.irma-international.org/article/the-prediction-of-workplace-turnover-using-machine-learning-technique/288055

Data Envelopment Analysis and Analytics Software for Optimizing Building Energy Efficiency
Zinovy Radovilsky, Pallavi Tanejaand Payal Sahay (2022). *International Journal of Business Analytics (pp. 1-17).*
www.irma-international.org/article/data-envelopment-analysis-and-analytics-software-for-optimizing-building-energy-efficiency/290404